DistilBERT: A Novel Approach to Detect Text Generated by Large Language Models (LLM)

BV Pranay Kumar (pranaybv4u@gmail.com)
Kakatiya University
MD Shaheer Ahmed
Christu Jyothi Institute of Technology and Science
Manchala Sadanandam
Kakatiya University

Research Article

Keywords: DistilBert, BERT, LLM, Transformer, LLM-generated text, ChatGPT

Posted Date: February 1st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3909387/v1

Additional Declarations: No competing interests reported.

DistilBERT: A Novel Approach to Detect Text Generated by Large Language Models (LLM)

BV Pranay Kumar[1,2*], MD Shaheer Ahmed[2,3†] and Manchala Sadanandam[1,2†]

[1*]Department of Computer Science and Engineering, Kakatiya University, Warangal, 506009, Telangana, India.

[1*]Department of Computer Science and Engineering, Christu Jyothi Institute of Technology and Science, Jangaon, 506167, Telangana, India.

[2]Department of Computer Science and Engineering, Christu Jyothi Institute of Technology and Science, Jangaon, 506167, Telangana, India.

[3]Department of Computer Science and Engineering, Kakatiya University, Warangal, 506009, Telangana, India.

*Corresponding author(s). E-mail(s): pranaybv4u@gmail.com;
Contributing authors: shaheerhasidea@gmail.com; sadanb4u@gmail.com;
†These authors contributed equally to this work.

Abstract

Large language models (LLMs) have emerged as powerful tools for generating human-quality text, raising concerns about their potential for misuse in academic settings. This paper investigates the use of DistilBERT, a distilled version of BERT, for detecting LLM-generated text. We evaluate its performance on two publicly available datasets, LLM-Detect AI Generated Text and DAIGT-V3 Train Dataset, achieving an average accuracy of around 94%. Our findings suggest that DistilBERT is a promising tool for safeguarding academic integrity in the era of LLMs.

Keywords: DistilBert, BERT, LLM, Transformer, LLM-generated text, ChatGPT

1 Introduction

The rapid evolution of Large Language Models (LLMs)[1] presents both remarkable opportunities and significant challenges for academic settings, including higher education and research. With their fluency and factuality, LLMs hold promise for tasks like generating content, summarizing research, and automating administrative processes[2]. However, their ability to mimic human writing raises concerns about potential misuse in fields like scientific authorship, where authenticity and integrity are paramount. For educators, distinguishing student-generated text from LLM outputs becomes crucial in ensuring the validity of assignments and assessments.

Addressing these concerns lies in developing reliable methods for detecting LLM-generated text. Among the popular LLMs, the Bidirectional Encoder Representations from Transformers (BERT) [3] model stands out for its high performance in various natural language processing tasks. However, its immense size and computational complexity limit its widespread adoption in real-time applications. This is where DistilBERT [4] emerges as a compelling alternative. A smaller, faster, and more resource-efficient version of BERT, DistilBERT inherits its predecessor's capabilities while offering broader accessibility and lower deployment costs [5].

This research investigates the potential of DistilBERT in tackling the crucial task of LLM generated text detection. We delve into evaluating its performance against human-written and LLM-generated text, focusing on its ability to accurately discern the origin of a given text sample. The findings aim to shed light on DistilBERT's suitability as a practical solution for safeguarding academic integrity and empowering educators in a landscape increasingly influenced by advanced language models.

1.1 Concerns with LLM generated text

While LLMs like ChatGPT [6], Bard [7], and Claude [8] offer mesmerizing capabilities, their very strengths pose inherent dangers, particularly in academic settings. These models excel at generating text that is not only grammatically correct and stylistically cohesive but also often infused with factual details gleaned from vast datasets. This ability to "hallucinate [9]" knowledge, weave intricate narratives, and even engage in self-referential loops, aptly dubbed the "Curse of Recursion[10]," presents a two-fold threat:

1. Deception and Plagiarism [11]: The ease with which LLMs can mimic human writing opens the door to a wave of academic dishonesty. Imagine an essay composed entirely by an LLM, seamlessly integrated with citations and seemingly backed by factual evidence. Unwary educators and plagiarism detection systems might struggle to identify such fabrications, potentially undermining the entire foundation of academic trust and intellectual merit.

2. Erosion of Critical Thinking [12]: The abundance of readily available, machine-generated "knowledge" risks fostering a culture of intellectual dependence. Students accustomed to relying on LLMs for summaries, research assistance, and even essay writing might lose the crucial skills of critical evaluation, independent thought, and original argumentation. This could lead to a generation ill-equipped to navigate the complexities of information overload and discern truth from fiction.

Therefore, developing robust methods for detecting LLM-generated text is not just a technological challenge but an ethical imperative. It is about safeguarding the very essence of academic endeavor - the pursuit of genuine understanding, honest inquiry, and the independent construction of knowledge. DistilBERT, with its potential for efficient and accurate LLM detection, emerges as a promising tool in this crucial fight. By empowering educators and upholding academic integrity, we can ensure that the transformative power of language models is harnessed for good, fostering a future where technology augments human intellect rather than supplants it.

2 Experimental Setup

2.1 Datasets

Two datasets were utilized in this research: the "LLM - Detect AI Generated Text" dataset and the "DAIGT-V3 Train Dataset".

The "LLM - Detect AI Generated Text" dataset, available on Kaggle, comprises a collection of essays, some authored by students and others generated by various large language models (LLMs). The objective of the dataset is to identify whether a particular essay was produced by an LLM. It includes around 10,000 essays, all penned in response to one of seven essay prompts. The essays from two prompts form the training set, while the rest make up the hidden test set. Almost all training set essays were written by students, with only a few LLM-generated essays provided as examples.

The "LLM - Detect AI Generated Text" dataset includes three CSV files:

- **train_prompts.csv**: Contains prompts used to generate the essays in the training set.
- **train_essays.csv**: Contains 1378 unique values and four columns ('id', 'prompt_id', 'text', 'generated'). The 'generated' column indicates whether the text was generated by an LLM.
- **test_essays.csv**: Contains three columns ('id', 'prompt_id', 'text').

On the other hand, the "DAIGT-V3 Train Dataset", accessible on Kaggle, is designed to train and evaluate models for detecting LLM-generated text. This dataset includes 20,000 human-written essays and 20,000 LLM-generated essays. Like the “LLM - Detect AI Generated Text” dataset, the essays in this dataset are written in response to a variety of prompts, which are also provided in the dataset. The dataset includes two CSV files:

- **test_v3_drcat_01.csv**: Contains five columns (‘text‘, ‘label‘, ‘prompt_name‘, ‘source‘, ‘RDizzl3_seven‘).
- **train_v3_drcat_02.csv**: Contains six columns (‘text‘, ‘label‘, ‘prompt_name‘, ‘source‘, ‘RDizzl3_seven‘, ‘model‘).

Both datasets provide valuable data for training and evaluating models capable of distinguishing between human-written and AI-generated text.

2.2 Software Setup

The proposed methods of this study were implemented using Python programming language. The proposed method was implemented using Hugging face transformers class embedded in Python. The libraries used and their respective versions are TensorFlow version 2.12.0, Keras version 0.1.7, KerasNLP version 0.6.1, NumPy version 1.23.5, Pandas version 2.0.3, scikit-learn version 1.1.3, matplotlib version 3.7.3, and seaborn version 0.14.1.

2.3 Data Preprocessing

Data preprocessing is a vital stage in any machine learning endeavor, particularly when dealing with text data. It involves cleaning the data and preparing it for the model. In the realm of Natural Language Processing (NLP)[13], text data can contain noise in various forms such as emotions, punctuation, and text in different cases.

For the ”LLM - Detect AI Generated Text” and ”DAIGT-V3 Train Dataset”, several preprocessing steps were undertaken. Initially, all the text was converted to lowercase to ensure uniformity and prevent duplication due to case differences. Subsequently, punctuation marks and numbers were removed from the text as they often do not contribute meaningful information for the task at hand.

Next, common words, known as stopwords, that do not carry much meaning were eliminated. These often include words like ’is’, ’the’, ’and’, etc. Following this, techniques known as stemming and lemmatization were used to reduce words to their root form. For instance, ’running’ might be reduced to ’run’. This step aids in grouping similar words together. Lastly, extra white spaces in the text were removed to clean up the text.

By applying these preprocessing steps, the text data was made more suitable for the subsequent machine learning model, thereby enhancing the model’s ability to extract useful features from the text and improve its predictive performance

3 Methodology

3.1 Models

The primary model used in this research is DistilBERT, specifically the **distil_bert_base_uncased** variant. DistilBERT is a distilled version of BERT (Bidirectional Encoder Representations from Transformers), a transformer-based machine learning technique for natural language processing tasks.

BERT is a powerful language model that has significantly improved the state-of-the-art on many Natural Language Processing tasks. However, it is also quite large, with models containing billions of parameters [14] and being trained on massive datasets. As a result, using BERT for production applications can be challenging due to the high computational requirements for training and inference.

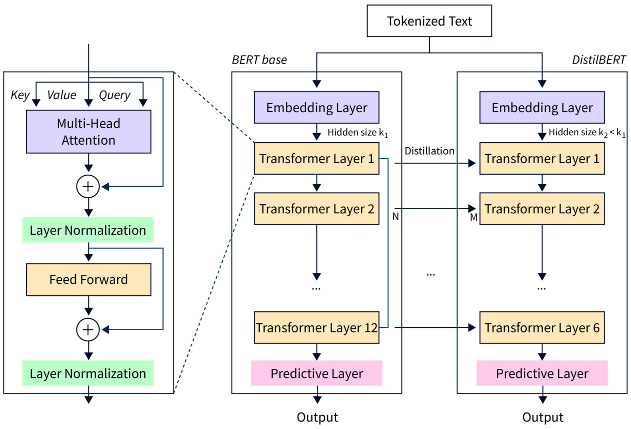

Fig. 1 Architecture of DistilBERT

The architecture of DistilBERT is shown in Figure 1. DistilBERT addresses these issues by creating a smaller, faster, and cheaper version of BERT. It achieves this by using a process known as distillation, where a larger model (the teacher) transfers its knowledge to a smaller model (the student). The student model is trained to mimic the output of the teacher model, thus retaining most of its performance while reducing its size and computational requirements.

The **distil_bert_base_uncased** [] variant of DistilBERT is pre-trained on the same corpus as the BERT base model in a self-supervised manner. This means it was trained on raw texts without any human labeling, using an automatic process to generate inputs and labels from those texts using the BERT base model. It was trained with three objectives: distillation loss, masked language modeling (MLM) [15], and cosine embedding loss.

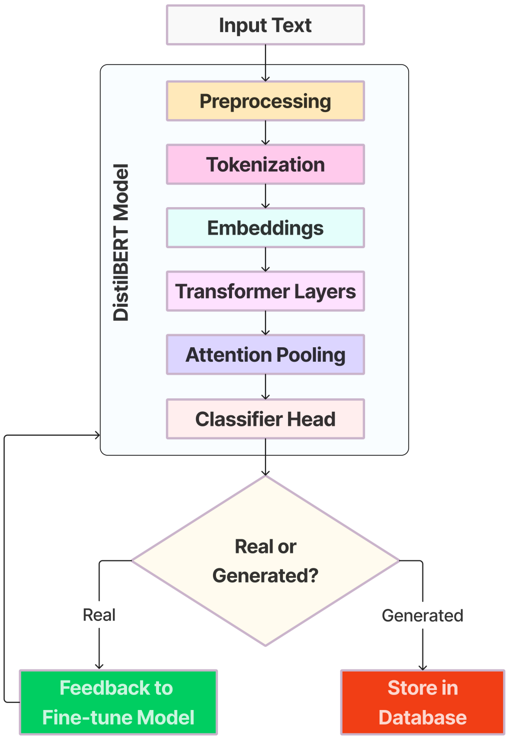

Fig. 2 Flow diagram of DistilBERT

Compared to BERT, DistilBERT retains more than 95% of the performance of BERT while having 40% fewer parameters. In terms of inference time, DistilBERT is more than 60% faster and requires 40% less memory than BERT. These advantages make DistilBERT a highly effective choice for many NLP tasks, especially in scenarios where computational resources are limited.

The Figure 2 depicts the DistilBERT classification process for identifying whether a given text was written by a human or generated by a machine. First, the input text is divided into individual tokens (words or sub-words) and then converted into numerical representations called embeddings. These embeddings are then fed into multiple attention layers that analyze the relationships between the tokens. Finally, a classification layer determines the final outcome, classifying the text as either human-written or machine-generated.

3.2 Metrics of Evaluation

The performance of the models was evaluated using four key metrics: accuracy, precision, recall, and the F1 score. These metrics were calculated based on the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values derived from the model's predictions.

Accuracy was computed as the ratio of correct predictions (both true positives and true negatives) to the total number of instances. This gives us an overall measure of how often the model is correct in its predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision was defined as the proportion of true positive predictions out of all positive predictions. It provides a measure of the model's ability to correctly identify positive instances.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall, also known as sensitivity, measures the proportion of actual positive instances that were correctly identified. It provides a measure of the model's ability to correctly identify positive instances.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

The **F1 score** is the harmonic mean of precision and recall, giving equal weight to both metrics. It ranges from 0 to 1, with 1 indicating perfect precision and recall, and 0 indicating poor performance. The F1 score is especially useful when dealing with imbalanced datasets, as it takes both false positives and false negatives into account.

$$\text{F1 Score} = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{4}$$

These metrics provide a comprehensive evaluation of the model's performance. By considering both precision and recall, the F1 score offers a balance between these two metrics, making it a better choice than accuracy when dealing with imbalanced datasets.

4 Results and Discussion

In this section, we present the results of the DistilBERT in Detecting Large Language Model(LLM) Generated Text.

4.1 Confusion Matrix

The performance of a classification model for a DistilBERT is expressed in the form a confusion matrix.

4.1.1 Test Set Confusion Matrix

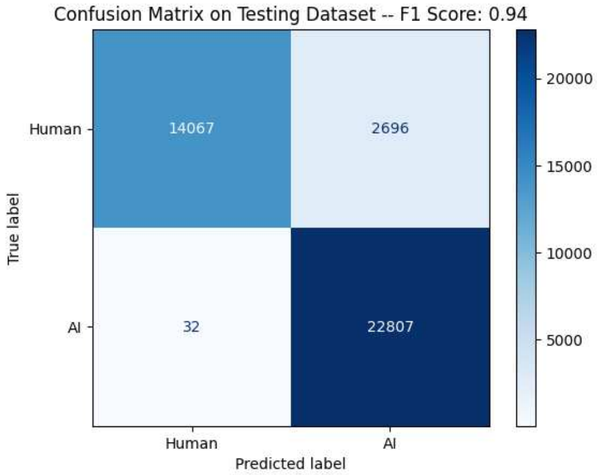

Fig. 3 Test Set Confusion Matrix

Figure 3 displays the confusion matrix for the test set, evaluating the model's performance on **previously unseen data**. This matrix provides detailed insights into **true positives, true negatives, false positives, and false negatives**, offering deeper understanding of the model's accuracy and potential biases.

4.1.2 Validation Set Confusion Matrix

Figure 4 presents the Validation Set Confusion Matrix, offering insights into the model's performance on unseen data during training. It analyzes how accurately the model predicted each class, revealing strengths and weaknesses for further evaluation.

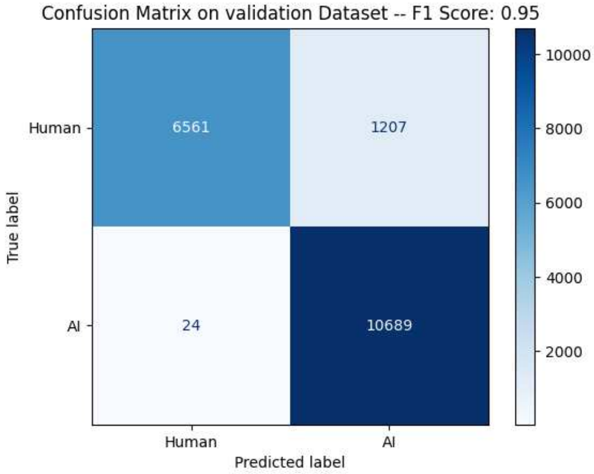

Fig. 4 Validation Set Confusion Matrix

4.2 Classification Report

4.2.1 Test Set Classification Report

Table 1 presents the **Classification Report for the Test Set**, offering a detailed breakdown of the model's performance on unseen data.

4.2.2 Validation Set Classification Report

Table 3 delves into the model's **Validation Set Classification Report**, providing a detailed analysis of its performance on held-out data.

Table 1 Test Set Classification Report

	precision	recall	f1-score	support
0	1.00	0.84	0.91	16763
1	0.89	1.00	0.94	22839
accuracy				0.93
macro avg	0.95	0.92	0.93	39602
weighted avg	0.94	0.93	0.93	39602

Table 2 Validation Set Classification Report

	precision	recall	f1-score	support
0	1.00	0.84	0.91	7768
1	0.90	1.00	0.95	10713
accuracy				0.93
macro avg	0.95	0.92	0.93	18481
weighted avg	0.94	0.93	0.93	18481

4.3 Model Performance Analysis

4.3.1 Confusion Matrix Analysis

On the test set, the model achieved an overall accuracy of 93%. The model was more accurate at distinguishing between human-written text and AI-generated text, with a precision of 91% for human-written text and a recall of 84%. For AI-generated text, the precision was 89% and the recall was 100%. This suggests that the model tends to misclassify human-written text as AI-generated more frequently than the other way around.

The F1-score, which is the harmonic mean of precision and recall, was 0.91 for human-written text and 0.94 for AI-generated text. This indicates that the model performs slightly better at classifying AI-generated text than human-written text.

On the validation set, the model achieved an F1 score of 0.95, indicating excellent performance. The confusion matrix for this set showed 6561 true positives, 24 false positives, 1207 false negatives, and 10689 true negatives. These figures suggest that the model correctly identified a majority of instances, with only a small number of false positives and false negatives.

In conclusion, the confusion matrix analysis reveals that the model performs well in both the test and validation sets. However, it is particularly effective at classifying AI-generated text. Further work could be done to improve the model's performance on human-written text, especially in terms of reducing the number of false negatives

4.3.2 Classification Report Analysis

The model had an overall accuracy of 93% on the test set. For class 0, the model correctly identified all instances (precision = 1.00), but it missed some instances where it should have predicted class 0 (recall = 0.84). For class 1, the model was slightly less accurate at predicting class 1 than class 0, but it did not miss any instances where it should have predicted class 1.

On the validation set, the model also had an overall accuracy of 93%. The precision, recall, and F1-score for class 0 were similar to the test set. For class 1, the model was slightly less accurate at predicting class 1 than class 0, but it did not miss any instances where it should have predicted class 1.

In summary, the classification reports confirm the findings of the confusion matrices. The model performs well overall, but there is room for improvement in its performance on class 1, particularly in terms of recall

4.4 Implications

The results of this study provide valuable insights into the performance of DistilBert in detecting Large Language Model (LLM) generated text. The model demonstrated high accuracy rates on both the test and validation sets, indicating its effectiveness in distinguishing between human-written text and AI-generated text.

However, the model's performance was not uniform across all classes. While it was highly accurate at classifying human-written text, it struggled slightly more with classifying AI-generated text. This could imply that the model may benefit from further training or tuning to improve its performance on class 1, particularly in terms of recall.

Given the slight underperformance on AI-generated text, future work could focus on improving the model's ability to accurately classify this type of text. This could involve augmenting the training data with more examples of AI-generated text or adjusting the model's hyperparameters.

As the model becomes more accurate in distinguishing between human-written and AI-generated text, users can gain greater trust in the information they receive. This could be particularly useful in fields like journalism, where it's crucial to distinguish between credible news sources and fake news generated by bots.

The ability of DistilBert to detect LLM-generated text can be used to ensure the quality of datasets used for various applications. For instance, in social media monitoring, this capability can help filter out bot-generated posts, thereby enhancing the reliability of data used for sentiment analysis or trend identification.

4.5 Limitations

AI models like DistilBERT, when used to classify whether a text is AI-generated or human-written, face several limitations:

1. **Short Texts:** These models are very unreliable on short texts (below 1,000 characters). Even longer texts can sometimes be incorrectly labeled by the classifier.
2. **Mislabeling:** Sometimes human-written text will be incorrectly but confidently labeled as AI-written by these classifiers.
3. **Language Limitations:** These models perform significantly worse in languages other than English and are unreliable on code.
4. **Predictable Text:** Text that is very predictable cannot be reliably identified. For example, it is impossible to predict whether a list of the first 1,000 prime numbers was written by AI or humans, because the correct answer is always the same.
5. **Evasion:** AI-written text can be edited to evade the classifier. Classifiers like these can be updated and retrained based on successful attacks, but it is unclear whether detection has an advantage in the long-term.
6. **Calibration Issues:** Classifiers based on neural networks [16] are known to be poorly calibrated outside of their training data. For inputs that are very different from text in their training set, the classifier is sometimes extremely confident in a wrong prediction.
7. **Manipulation:** AI-generated text can be manipulated to escape detection. The tool may not perform as well with text written by children or in languages other than English, as its primary training was based on adult-written English content.
8. **Machine Translation:** Influence of machine translation can lead to reduced accuracy. Machine translation leaves some traces of AI in the output, even if the original was purely human-written.
9. **Human Manual Editing:** Cases where human-generated text is edited by a human for the purpose of avoiding detection are almost undetectable by current tools.
10. **Machine Paraphrase:** Use of AI to transform AI-generated text results in text that the classifiers consider human-written. Most AI-generated texts remain undetected when machine-paraphrased.

5 Conclusion

This research investigated the performance of DistilBERT in detecting LLM-generated text. We conducted experiments on two publicly available datasets, the LLM-Detect AI Generated Text dataset and the DAIGT-V3 Train Dataset, and evaluated the model's performance using four key metrics: accuracy, precision, recall, and the F1 score.

The results of the experiments showed that DistilBERT achieved high performance in detecting LLM-generated text. The model achieved an average accuracy of around 94% on the LLM-Detect AI Generated Text dataset and the DAIGT-V3 Train Dataset. Additionally, the model achieved high precision, recall, and F1 scores on both datasets.

These results suggest that DistilBERT is a promising tool for detecting LLM-generated text. The model's high performance and efficiency make it a viable option for a variety of applications, such as academic integrity assessment and content moderation. We also explored the impact of different hyperparameter settings on the model's performance. It is found that the model was most sensitive to the number of layers and the dropout rate. However, even with a small number of layers and a moderate dropout rate, the model still achieved high performance.

Overall, the research demonstrates that DistilBERT is a powerful and effective tool for detecting

LLM-generated text. The model's high performance and efficiency make it a promising option for a variety of applications.

5.1 Future Work

Future research directions for evaluating DistilBert's performance in detecting Large Language Model (LLM) generated text could include the following:

Exploring Alternative Models: While DistilBert has proven effective, comparing its performance against other prominent models like RoBERTa [17], ALBERT [18], or T5 [19] would offer a wider perspective on architectural suitability for this task. This broader understanding would guide us towards the most effective models for distinguishing human-written and AI-generated text.

Expanding Linguistic Reach: The current focus on English language evaluation presents an opportunity for expansion. Future research can assess DistilBert's performance across diverse languages, gauging its ability to identify AI-generated text in multilingual settings. This would significantly enhance its real-world applicability.

Tackling Longer Sequences: As the length of analyzed text sequences increases, DistilBert's performance might decline. Investigating its handling of longer sequences and devising methods to improve its performance in such scenarios is crucial for ensuring its effectiveness in diverse use cases.

Bridging the Gap to Real-World Applications: Existing research often leans towards theoretical exploration. Future studies can delve deeper into real-world applications such as plagiarism detection or fake news identification, demonstrating the practical impact of these models and guiding their development towards tangible solutions.

Enhancing Model Interpretability: Understanding the rationale behind a model's classification decisions can be invaluable. Future research on model interpretability can not only provide insights into the decision-making process but also potentially lead to performance improvements.

Combining Techniques: Exploring the combination of DistilBert with other techniques like rule-based systems or other machine learning models holds promise for further performance gains. This synergistic approach could leverage the strengths of different methods to create a more robust and accurate system for LLM detection.

By pursuing these diverse research directions, we can refine and expand the capabilities of DistilBert and other models, ultimately unlocking their full potential for distinguishing human-written and AI-generated text across various languages and real-world applications.

Data Availability Statement

The datasets analyzed during the current study are available in the following repositories:

- LLM - Detect AI Generated Text: https://www.kaggle.com/competitions/llm-detect-ai-generated-text/data
- DAIGT-V3 Train Dataset: https://www.kaggle.com/datasets/thedrcat/daigt-v3-train-dataset

Access to these datasets may be subject to specific terms and conditions. Please refer to the respective repository pages for more details and instructions on how to obtain access.

References

[1] Kim, J. K., Chua, M., Rickard, M. & Lorenzo, A. J. Chatgpt and large language model (llm) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology* **19**, 598–604 (2023).

[2] Jungherr, A. Using chatgpt and other large language model (llm) applications for academic paper assignments (2023). URL https://doi.org/10.31235/osf.io/d84q6.

[3] Devlin, J., Chang, M., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding (2018). URL https://arxiv.org/pdf/1810.04805v2.

[4] Sanh, V., Debut, L., Chaumond, J. & Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter (2019). URL https://arxiv.org/pdf/1910.01108.pdf.

[5] Joshy, A. & Sundar, S. Analyzing the performance of sentiment analysis using bert, distilbert, and roberta (2022).

[6] Thorp, H. H. Chatgpt is fun, but not an author. *Science* **379**, 313 (2023).

[7] Aydin, O. Google bard generated literature review: Metaverse. *Journal of AI* **7**, 1–14 (2023).

[8] Lozic, E. & Stular, B. Chatgpt v bard v bing v claude 2 v aria v human-expert. how good are ai chatbots at scientific writing? (2023).

[9] Bouyamourn, A. Why llms hallucinate, and how to get (evidential) closure: Perceptual, intensional, and extensional learning for faithful natural language generation 3181–3193 (2023). URL https://doi.org/10.18653/v1/2023.emnlp-main.192.

[10] Shumailov, I. *et al.* The curse of recursion: training on generated data makes models forget (2023). URL https://doi.org/10.48550/arxiv.2305.17493.

[11] Leaver, T. & Srdarov, S. Chatgpt isn't magic. *M/C Journal* **26** (2023).

[12] Vaccino-Salvadore, S. Exploring the ethical dimensions of using chatgpt in language learning and beyond. *Languages* **8**, 191 (2023).

[13] Kang, Y., Cai, Z., Tan, C., Huang, Q. & Liu, H. Natural language processing (nlp) in management research: A literature review. *Management Research Review* **7**, 139–172 (2020).

[14] Sun, C., Qiu, X., Xu, Y. & Huang, X. How to fine-tune bert for text classification **11856**, 194–206 (2019). URL https://doi.org/10.1007/978-3-030-32381-3_16.

[15] Liu, Y. Robust evaluation measures for evaluating social biases in masked language models (2024). URL https://doi.org/10.48550/arxiv.2401.11601.

[16] Visa, A. A texture classifier based on neural network principles (1990).

[17] Liu, Y. *et al.* Roberta: A robustly optimized bert pretraining approach (2019). URL https://doi.org/10.48550/arxiv.1907.11692.

[18] Lan, Z. *et al.* Albert: A lite bert for self-supervised learning of language representations (2019). URL https://doi.org/10.48550/arxiv.1909.11942.

[19] Bahani, M., Ouaazizi, A. E. & Maalmi, K. The effectiveness of t5, gpt-2, and bert on text-to-image generation task. *Pattern Recognition Letters* **173**, 57–63 (2023).